EndoRD-GS: Robust Deformable Endoscopic Scene Reconstruction via Gaussian Splatting

Bingchen Gao, Jun Zhou, Jing Zou, and Jing Qin, *Senior Member, IEEE*

***Abstract*— Real-time and realistic reconstruction of 3D dynamic surgical scenes from surgical videos is a novel and unique tool for surgical planning and intraoperative guidance. The 3D Gaussian splatting (GS), with its high rendering speed and reconstruction fidelity, has recently emerged as a promising technique for surgical scene reconstruction. However, existing GS-based methods still have two obvious shortcomings for realistic reconstruction. First, they largely struggle to capture localized yet intricate soft tissue deformations caused by complex instrument-tissue interactions. Second, they fail to model spatiotemporal coupling among Gaussian primitives for global adjustments during rapid perspective transformations, resulting in unstable reconstruction outputs. In this paper, we propose *EndoRD-GS*, an innovative approach that overcomes these two limitations through two core techniques: (1) periodic modulated Gaussian functions and (2) a new Biplane module. Specifically, our periodic modulated Gaussian functions incorporate meticulously designed modulations, significantly enhancing the representation of complex local tissue deformations. On the other hand, our Biplane module constructs spatiotemporal interactions among Gaussian primitives, enabling global adjustments and ensuring reliable scene reconstruction during rapid perspective transformations. Extensive experiments on three datasets demonstrate that our EndoRD-GS achieves superior performance in endoscopic scene reconstruction compared to state-of-the-art methods.**

***Index Terms*— Endoscopic scene reconstruction, Gaussian splatting, Periodic modulated Gaussian functions, Biplane module**

I. Introduction

ENDOSCOPIC procedures have become indispensable in modern surgery. However, the restricted field of view poses significant challenges for surgeons, who have to operate in confined spaces using elongated instruments without direct three-dimensional visual guidance [3], [4]. This constraint compels surgeons to navigate complex anatomical structures through two-dimensional endoscopic imaging, compromising their ability to accurately identify critical anatomical landmarks. To mitigate this constraint, one promising solution is to reconstruct a dynamic 3D environment of current surgical scene to facilitate 3D navigation. Such a 3D environment, once well reconstructed and deployed in operation room, is capable of enhancing 3D anatomical understanding of the current scene [5], tracking deformed organs [6], and underpinning autonomous robotic manipulations [7].

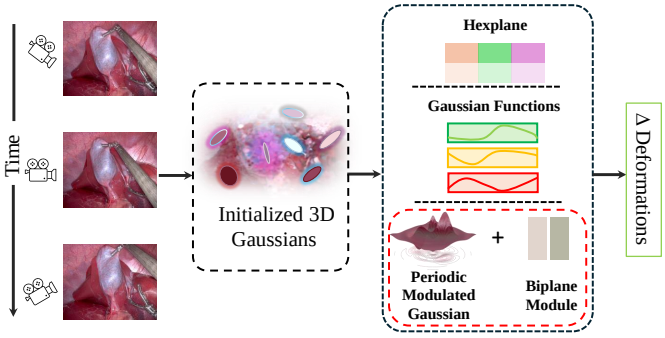


Fig. 1: Comparison of Gaussian primitive deformation modeling methods. Previous approaches employ HexPlane [1] or Gaussian basis functions [2] to model Gaussian primitive deformations. Our method, highlighted in the red box, introduces periodic modulated Gaussian functions and the Biplane module for modeling Gaussian primitive deformations.

However, non-rigid soft tissue deformation caused by instrument-tissue interaction increases difficulties in scene reconstruction, making precise representation of 3D endoscopic scenes a formidable challenge [8]. A lot of effort has been dedicated to surgical scene reconstruction in the past years. Early investigations proposed to reconstruct scenes through depth estimation [9], [10], point cloud fusion (in a SLAM-style framework) [3], [11], [12], and warp field integration [13]. However, depth estimation methods, operating on individual frames and highly sensitive to noise, fail to deliver a temporally coherent model of the operative field [14]. The SLAM-based methods struggle with rapid, non-rigid tissue motions and impose substantial computational overhead. The warp field integration approaches often introduce reconstruction artifacts. To the end, these techniques are still far from satisfactory to realistically or robustly capture dynamic changes inherent in surgical scenes. Recent advances have increasingly turned to deep learning techniques, particularly neural radiance fields

This work was supported in part by a General Research Fund of Hong Kong Research Grants Council (project no. 15218521), and in part by a Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under Shenzhen Municipal Science and Technology Innovation Commission (project no. SGDX20230821092359002).

Bingchen Gao, Jun Zhou, Jing Zou, and Jing Qin are with the Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, HKSAR, China. (e-mail: bingchen.gao@connect.polyu.hk; zachary-jun.zhou@connect.polyu.hk;zoujing.zou@polyu.edu.hk; harry.qin@polyu.edu.hk).

Bingchen Gao and Jing Qin is also with the PolyU-Qianhai Disruptive Technology and Innovation Research Centre (QHRC).

Bingchen Gao and Jun Zhou contributed equally.

Corresponding author: Jing Zou.

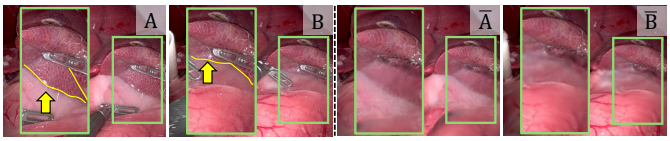
(NeRFs), for endoscopic scene reconstruction [15]–[18]. The EndoNeRF [16] pioneers the application of NeRFs by modeling surgical scenes as a canonical field with time-dependent displacement, while EndoSurf [17] incorporates a signed distance field (SDF) [19], [20] to explicitly constrain surface geometry. Although these NeRF-based methods demonstrate promising results, they represent the geometry of surgical scenes via implicit neural fields, which leads to protracted training time (order of hours) and suboptimal rendering speed, significantly limiting their clinical applicability.

To achieve fast rendering, some studies have adapted 3D Gaussian splatting (GS) [21], a recently developed 3D scene representation method, to endoscopic scene reconstruction. Most of these studies build on the foundational EndoGS [22], which employs HexPlane [1] to formulate 4D Gaussian primitives with time-dependent deformable parameters for representing dynamic surgical scenes [2], [23]–[26]. Although these approaches achieve notable rendering speed, their HexPlane-based 4D factorization for deformation modeling places a significant burden on a subsequent compact multilayer perceptron (MLP). This small MLP cannot adequately extract the decomposed plane features and hence fails to efficiently model deformation parameters for Gaussian primitives. To meet this challenge, Deform3DGS [2], the current state-of-the-art method, proposes to use learnable Gaussian basis functions to model deformations for each Gaussian primitive.

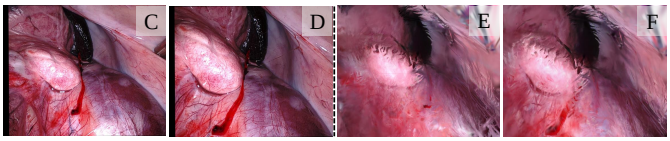
Despite its efficiency in modeling Gaussian primitive deformations, Deform3DGS faces some major challenges in realistically capturing dynamic tissue deformations and movements. **First**, as illustrated in Fig. 2a, the instrument-tissue interactions may vary significantly due to different tissue properties, surgical maneuvers, and surgeon experience levels, resulting in heterogeneous tissue motion patterns [27]–[30]. However, the smooth exponential decay property of Gaussian basis functions in Deform3DGS constrains their capability in generating rapid parameter changes necessary for simulating diverse tissue motion patterns. Consequently, the method struggles to capture fine-grained local tissue movements within short time intervals. **Second**, as depicted in Fig. 2b, endoscopic procedures often involve rapid perspective transformations, as surgeons require free camera movement to observe different tissues during interventions [31], [32]. These rapid transformations make it challenging to adjust Gaussian primitive parameters simultaneously, necessitating a comprehensive modeling of spatial and temporal relationships between primitives for global adjustments. Nevertheless, Deform3DGS applies Gaussian basis functions individually to each primitive; it lacks a mechanism to account for the inter-primitive relationships, limiting its capacity to model global deformation parameters under rapid perspective transformations.

To address these challenges, we propose EndoRD-GS, a novel approach for deformable endoscopic scene reconstruction. In response to the first challenge, we present a periodic modulated Gaussian function that incorporates learnable periodic modulations. This modulation facilitates temporal oscillations of the Gaussian functions and provides amplitude control, thereby enhancing the model's ability to capture complex local tissue deformations. To tackle the second challenge, we introduce the Biplane module, which models spatiotemporal coupling throughout the scene and enables a comprehensive global adjustment of Gaussian primitive deformation parameters. Specifically, our Biplane module employs a high-dimensional grid strategy to achieve global refinement of Gaussian primitives while maintaining computational efficiency during rapid perspective transformations. The main contributions of this work are summarized as follows:

1) We propose EndoRD-GS (Fig. 1), a novel approach for robust dynamic 3D endoscopic scene reconstruction.
2) We introduce the periodic modulated Gaussian functions for capturing local tissue deformation alongside a Biplane module for globally modeling the spatiotemporal interactions between Gaussian primitives during endoscopic camera movements.
3) Extensive experiments conducted on three representative datasets demonstrate that EndoRD-GS achieves state-of-the-art performance in endoscopic scene reconstruction.



(a) **Challenge 1: Local Tissue Deformation.** Figures A and B present ground truth tissue deformations under instrument-tissue interactions. In contrast, the corresponding renderings by the latest GS model [2] (Figures Ā and B̄) fail to capture local deformations, particularly subtle texture changes and edge shape variations.



(b) **Challenge 2: Global Adjustment for Gaussian Primitives During Perspective Transformations.** Figures C-D display rapid motions in camera position between consecutive frames. The latest GS model [2] fails to globally adjust Gaussian primitives during these rapid camera motions, resulting in noticeable artifacts in the renderings, as shown in Figures E-F.

Fig. 2: Comparative illustrations of challenges related to local tissue deformation patterns and rapid perspective transformations. The left section displays the ground truth, while the right section presents the corresponding rendering results.

II. Related Work

A. Traditional Scene Reconstruction Techniques

Traditional methods for reconstructing 3D structures rely on estimating camera poses and feature matching. One commonly used technique is the structure from motion (SfM), which has been extensively applied in reconstructing scenes, organs, and tissues using endoscopic images [33], [34]. This technique constructs 3D models by analyzing a series of two-dimensional images. However, SfM methods like COLMAP [35] often struggle to accurately estimate camera poses throughout the sequence, leading to unstable reconstruction results. Simultaneous localization and mapping (SLAM) is another technique employed in endoscopic scene reconstructions [36],

[37]. It builds maps of unknown environments while concurrently estimating the camera's position and orientation. SLAM-based approaches have been used to model both sparse and dense tissue surfaces. While DefSLAM [36] and SD-DefSLAM [37] are able to reconstruct deformable scenes, they deform surfaces from pre-defined reference templates. This deformable model assumes isometric deformations and fails to accurately capture the non-isometric deformations prevalent in complex surgical interactions, potentially leading to unnatural shape estimations. Moreover, their reliance on iterative SLAM optimization imposes considerable computational overhead, limiting their practical application in surgical scenarios.

B. Neural Radiance Fields for Scene Reconstruction

Neural radiance fields (NeRFs) [38] aim to learn a continuous representation of 3D scenes, enabling high-quality novel view synthesis. Unlike traditional SLAM/SfM techniques that typically learn discrete representations, NeRF continuously encodes the appearance and geometry of scenes implicitly. Various approaches have been developed to extend NeRF for dynamic scene reconstruction, including D-NeRF [39], RoDynRF [40], EndoNeRF [16], and EndoSurf [17]. Recent NeRF-based methods have been developed to address the unique challenges in endoscopic surgical scene reconstruction, including fish-eye distortion, vignetting effects, and non-uniform illumination. These methods propose various specialized solutions: Neural Fields for 3D Tracking [41] enables continuous spatiotemporal reconstruction from monocular endoscopic videos, while ForPlane [42] accelerates 4D reconstruction by separating static and dynamic scene components. BASED [43] advances camera pose-free reconstruction through bundle adjustment. LightNeuS [15] explicitly addresses illumination degradation in endoscopic environments, and Sun et al. [44] enhances reconstruction accuracy by integrating depth estimation with dynamic NeRF modeling. Although recent works [42], [45] have made progress in accelerating NeRF-based surgical scene reconstruction through efficient network architectures and data representations, these methods still suffer from slow rendering speed when applied to dynamic, high-fidelity surgical scene reconstruction scenarios, limiting their intraoperative applicability.

C. Gaussian Splatting for Scene Reconstruction

Unlike NeRF-based methods, which encode the scenes implicitly, 3D-GS employs 3D Gaussian primitives for explicit scene representation and utilizes a cuda-accelerated tile-based rasterizer for real-time 3D rendering. While the original work focuses on static scenes, recent extensions adapt the original static scene representation to accommodate temporal changes and object deformations in videos [46]–[50]. Furthermore, several specialized methods have been developed in endoscopic scene reconstruction. EndoSparse [24] tackles limited view scenarios by incorporating prior knowledge and sparse view synthesis, while LGS [25] enhances computational efficiency by optimizing Gaussian selection and simplifying attribute representations. For dynamic scene reconstruction, several approaches extend Gaussian Splatting to 4D modeling. EndoGaussian [23] and EndoGS [22] employ lightweight MLPs to model Gaussian primitive deformations, while Endo-4DGS [51] further integrates depth estimation into the 4D Gaussian splatting framework. Further advancing the field, in terms of modeling dynamic changes of Gaussian primitives, Deform3DGS [2] adopts learnable Gaussian basis functions to model the temporal deformation of Gaussian primitives. Due to its flexibility and adaptability, we adopt Deform3DGS [2] as our baseline model. These methods collectively represent significant progress in endoscopic scene reconstruction yet leave room for improvement in modeling and adapting Gaussian primitives during surgery procedures.

III. Method

A. Preliminaries of 3D Gaussian Splatting

The 3D Gaussian splatting (GS) [21] represents a static 3D scene as a collection of unstructured 3D Gaussian primitives characterized by their spatial properties and appearance attributes. Specifically, the 3D Gaussians is defined by:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{1}$$

where $\mathbf{\Sigma}$ is the 3D covariance matrix and μ is the mean of the Gaussian primitive. These 3D Gaussians are projected onto 2D space with covariance $\mathbf{\Sigma}' = \mathbf{JW\Sigma W}^T\mathbf{J}^T$, where $\mathbf{\Sigma}'$ is the covariance matrix in the 2D plane, $\mathbf{W}$ is the view transformation and $\mathbf{J}$ is the projective transformation's Jacobian. To ensure positive semi-definiteness, the $\mathbf{\Sigma}$ is parameterized as $\mathbf{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$ with scale $\mathbf{S}$ and rotation $\mathbf{R}$. The final color C of pixel p is rendered using point-based volume rendering:

$$C(p) = \sum_{k\in\mathcal{N}} c_k\alpha_k \prod_{j=1}^{k-1}(1-\alpha_j), \tag{2}$$

Where k denotes the index of the Gaussian being evaluated, j represents the indices of all Gaussians closer to the camera than k, α_k is the final opacity value, c_k is the Gaussian color (modeled using spherical harmonics for view dependency), and $\prod_{j=1}^{k-1}(1-\alpha_j)$ represents the accumulated transparency from all closer Gaussians with indices j.

B. EndoRD-GS Pipeline

The architecture of the proposed EndoRD-GS is illustrated in Fig. 3. Let (X, Y) represent the joint space of images and camera poses, where $X = \{x_i\}_{i=1}^{H}$ corresponds to H video frames, and $Y = \{y_i\}_{i=1}^{H}$ denotes the associated camera poses. Each camera pose y_i includes the intrinsic matrix K and the extrinsic matrix T_i. Since a dense distribution of Gaussian primitives aids in reconstructing highly deformed regions in surgical scenes, we initialize the Gaussian point cloud using the motion-aware point fusion (MAPF) scheme [2]. Specifically, let $G = \{g_p^q\}_{p=1,q=0}^{P,2}$ denote the Gaussian primitives cloud at stage q, $q \in \{0, 1, 2\}$, where p is the index and P denotes the total number of the Gaussian primitives. Each initialized primitive g_p^0 is characterized by center $\mu_p^0 \in \mathbb{R}^3$, rotation $\mathbf{r}_p^0 \in \mathbb{R}^4$, scale $\mathbf{s}_p^0 \in \mathbb{R}^3$, opacity α_p^0, and spherical harmonic coefficients $\mathbf{c}_p^0$. After Gaussian primitive initialization, we apply the periodic modulated Gaussian function $\tilde{b}^*(t;\Phi)$ to transform each Gaussian primitive g_p^0 into

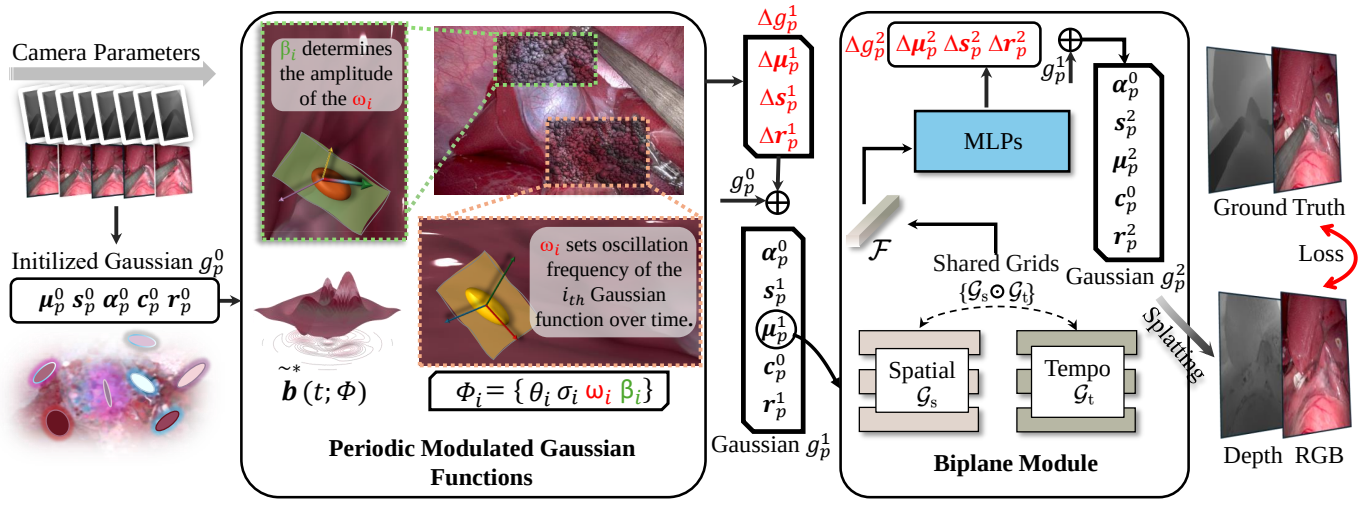


Fig. 3: The pipeline of our EndoRD-GS approach proceeds as follows: Given initialized Gaussian primitives g_p^0 from endoscopic scene videos, we first utilize periodic modulated Gaussian functions to capture deformations, transforming g_p^0 to g_p^1. Then, we integrate the Biplane module to establish spatiotemporal relationships between g_p^1, globally transforming g_p^1 to g_p^2.

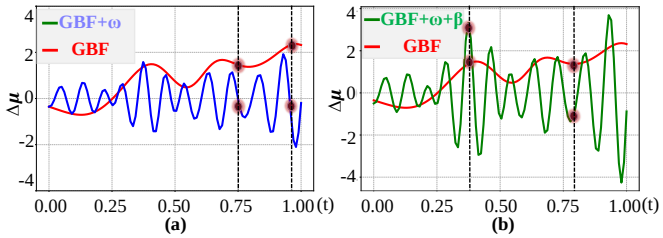


Fig. 4: Comparisons between the standard Gaussian Basis Function (GBF) and the modified version with frequency ω and amplitude β parameters. (a) The ω introduces more oscillations, enabling varying deformation patterns. (b) The β adjusts the scale of oscillations, allowing for a broader range of deformation magnitudes.

g_p^1 (in Sec. III-C). Next, we implement the Biplane module to dynamically modify each Gaussian primitive from g_p^1 to g_p^2, adapting to endoscopic perspective transformations (in Sec. III-D). Through integrating these techniques, EndoRD-GS optimizes the GS-model f_Ψ, parameterized by Ψ, by minimizing the rendering loss:

$$\min_{f_\Psi} \frac{1}{H} \sum_{i=1}^{H} L(f_\Psi(x_i, y_i), x_i), \tag{3}$$

where x_i represents the i-th video frame, y_i is the corresponding camera pose, H is the frame number, and L is the rendering loss function comparing the model's output $f_\Psi(x_i, y_i)$ with the ground truth frame x_i. Through this optimization process, the EndoRD-GS approach yields the trained GS model f_{Ψ^*} capable of reconstructing endoscopic scenes.

C. Periodic Modulated Gaussian Functions

To simulate local tissue deformations, Deform3DGS [2] utilizes Gaussian basis functions to model Gaussian primitive deformations. Specifically, this Gaussian basis function $\tilde{b}$ model the temporal evolution of deformation parameters for each Gaussian primitive. This basis function is defined by learnable parameters of temporal center θ and width σ:

$$\tilde{b}(t;\theta,\sigma) = \exp\Big(-\frac{1}{2\sigma^2}(t-\theta)^2\Big). \tag{4}$$

By adjusting θ and σ, the $\tilde{b}$ describes how the deformation parameters of Gaussian primitives evolve over time t. However, the inherent unimodal and smoothly decaying profile of this $\tilde{b}$, which peaks at its temporal center θ, restricts its capacity to capture complex temporal dynamics. In other words, it struggles to adequately represent the non-linear evolution patterns of deformation parameters for simulating intricate local tissue motions. As instantiated in Fig. 4, the Gaussian basis function (red line) exhibits monotonically increasing behavior in the interval [0.75, 1], limiting its capacity to represent diverse deformation parameter variations within this temporal window. To address this constraint, we leverage periodic functions, mainly cosine functions, which inherently repeat at regular intervals. The critical insight is that higher-frequency periodic functions complete more oscillation cycles within a given temporal window, enabling the representation of diverse deformation parameter variations. Meanwhile, to ensure that these diverse parameter variations represent accurate reconstruction, we argue that this periodic modulation enhances the learnable capacity in adaptively capturing the diverse Gaussian variations across different timestamps, thereby precisely simulating continuous variations in surgical procedures.

Specifically, we enhance the current Gaussian basis function $\tilde{b}$ by introducing two learnable parameters ω, β that control the frequency and amplitude of oscillations, respectively. The frequency parameter ω governs the rate of oscillation in the temporal dimension, representing periodic changes. By adjusting the frequency parameter ω in $\cos(\omega t)$, we can effectively modulate the oscillation frequency to capture diverse deformation parameter variations within specific temporal intervals. The learnable amplitude parameter β controls the oscillation

amplitude, allowing function $\tilde{b}$ to autonomously generate deformation parameters across a broad spectrum of values at specific timestamps, thereby simulating various magnitudes of tissue deformation. Formally, let $\Phi = \{\theta, \sigma, \omega, \beta\}$ be the set of the learnable parameters, the periodic modulated Gaussian function $\tilde{b}^*$ that describe the deformations along the temporal axis for each primitive g_p^0 is formulated as follows:

$$\tilde{b}^*(t;\Phi) = \exp\left(-\frac{1}{2\sigma^2}(t-\theta)^2\right)\cdot\beta\cdot\cos(\omega t), \tag{5}$$

where $\tilde{b}^*$ can be utilized to construct comprehensive Gaussian function curves through superposition. Taking the modulation of positional parameters as an example, we formulate the corresponding Gaussian function curve as:

$$\psi_\mu(t;\Phi^\mu) = \sum_{i=1}^{B}\tilde{b}(t;\theta_i^\mu,\sigma_i^\mu,\omega_i^\mu,\beta_i^\mu), \tag{6}$$

where ψ_μ enables precise computation of positional displacement $\Delta\mu_t$ at given temporal instance t. As demonstrated in Fig. 4, the frequency-modulated Gaussian basis function (blue line) exhibits multiple patterns of shape variations compared to the Gaussian basis function (red line), enabling the representation of varied patterns of positional displacement $\Delta\mu$ within the interval [0.75, 1]. Furthermore, with the addition of the amplitude parameter β (green line), the Gaussian function curves demonstrate an enhanced capacity to represent a broader range of positional displacement $\Delta\mu$. Under the same operation, the periodically modulated Gaussian function $\tilde{b}^*$ predicts the deformation parameter set Δg_p^1, comprising position ($\Delta\mu_p^1$), rotation ($\Delta\mathbf{r}_p^1$), and scale ($\Delta\mathbf{s}_p^1$) parameters. The Gaussian primitive g_p^1 are then obtained through linear combination of Δg_p^1 with g_p^0.

D. Biplane Module

Besides local intricate deformations, rapid camera movements lead to significant object shifts in the viewpoint in endoscopic procedures, necessitating further global adjustments of the Gaussian primitives. However, relying solely on Gaussian functions lacks the capability for global primitive adjustments due to insufficient modeling of the spatiotemporal coupling between Gaussian primitives. To model this coupling, we propose a Biplane module, illustrated in Fig. 5, that adopts a grid-based representation inspired by HexPlane [1]. Considering the camera's orientation, we consolidate the predominantly x-y planar spatial information into a shared 2D x-y plane grid while handling temporal dynamics through a separate shared 2D z-t plane grid. Specifically, let $\varphi = \{\mu_{p_x}^1, \mu_{p_y}^1, \mu_{p_z}^1, t\} \in \mathbb{R}^4$ denote the center of a Gaussian primitive in spacetime, where $\{\mu_{p_x}^1, \mu_{p_y}^1, \mu_{p_z}^1\}$ are the spatial coordinates and t is the temporal coordinate. We introduce two shared grids: the spatial and temporal shared grids, denoted as $\mathcal{G}_\mathrm{s}$ and $\mathcal{G}_\mathrm{t}$, which are initialized as learnable tensors over normalized domains[-1,1]:

$$\begin{cases}\mathcal{G}_\mathrm{s} = \left\{\mathbf{g}_{d,i,j} \mid \mathbf{g}_{d,i,j} \in \mathbb{R}^D, d\in[D], i\in[N_x], j\in[N_y]\right\}\\ \mathcal{G}\mathrm{t} = \left\{\mathbf{h}_{d,k,l} \mid \mathbf{h}_{d,k,l} \in \mathbb{R}^D, d\in[D], k\in[N_z], l\in[N_t]\right\},\end{cases} \tag{7}$$

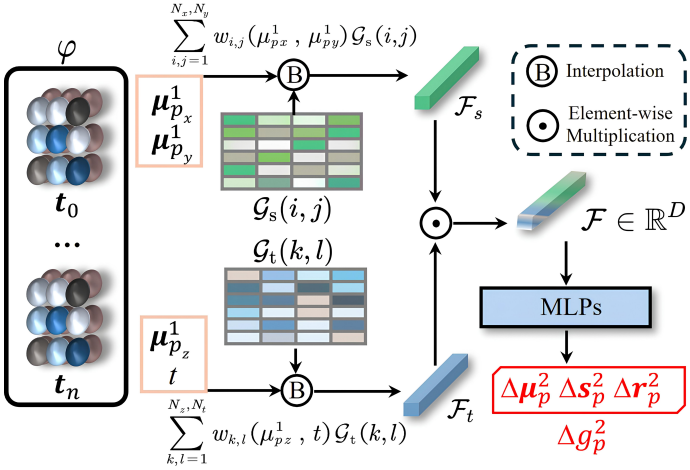


Fig. 5: Illustration of the Biplane module. Given the input set φ containing the position μ_p^1 of Gaussian primitives g_p^1 and time coordinate t, we utilize the spatial $\mathcal{G}$s and temporal $\mathcal{G}$t grids to predict the deformation parameters Δg_p^2.

where indices $i \in [N_x]$, $j \in [N_y]$, $k \in [N_z]$ and $l \in [N_t]$ correspond to the spatial and temporal resolutions, while $d \in [D]$ indexes the feature dimensions. After initializing the shared grids, we obtain the final voxel feature representation $\mathcal{F} \in \mathbb{R}^D$ by interpolating from both the spatial and temporal shared grids and fusing the results as follows:

$$\mathcal{F} = \left(\sum_{i,j=1}^{N_x,N_y} w_{i,j}(\mu_{p_x}^1,\mu_{p_y}^1)\mathcal{G}_\mathrm{s}(i,j)\right)\odot\left(\sum_{k,l=1}^{N_z,N_t} w_{k,l}(\mu_{p_z}^1,t)\mathcal{G}_\mathrm{t}(k,l)\right), \tag{8}$$

where $w_{i,j}(\mu_{p_x}^1, \mu_{p_y}^1)$ and $w_{k,l}(\mu_{p_z}^1, t)$ are the bilinear interpolation weights at spatial coordinates $(\mu_{p_{xyz}}^1)$ and temporal coordinate t, respectively. $\mathcal{G}_\mathrm{s}(i,j)$ and $\mathcal{G}_\mathrm{t}(k,l)$ are the grid values at the corresponding indices, and $\odot$ denotes element-wise multiplication. Furthermore, by leveraging shared grids, we significantly decrease the number of interpolation operations and element-wise multiplications. Compared to HexPlane, the computational burden per input primitive is reduced from $N_p \cdot k_i D + (N_p - 1)\cdot k_m D$ to $2\cdot k_i D + k_m D$, where $N_p = 6$ is the number of planes in the HexPlane method, k_i and k_m are constants representing the computational cost per feature dimension for interpolation and element-wise multiplication.

The Biplane module effectively captures spatiotemporal interactions while significantly reducing computational cost by using shared grids for interpolation and one single fusion operation. Next, we employ three tiny MLPs to decode the deformation parameter set Δg_p^2, comprising position ($\Delta\mu_p^2$), rotation ($\Delta\mathbf{r}_p^2$), and scale ($\Delta\mathbf{s}_p^2$) parameters, from the voxel feature $\mathcal{F}$. We finally obtain the final deformed Gaussian primitive g_p^2 by linearly combining Δg_p^2 with g_p^1.

E. Optimization and Inference

By implementing the periodic modulated Gaussian functions and the Biplane module, Our EndoRD-GS approach modulates the deformed Gaussian primitives g_p^2 to reconstruct the endoscopic scene. To optimize the EndoRD-GS, given a tissue mask M, we train the reconstruction model Ψ by

supervising the rendered images and depths splatted via g_p^2 against ground-truth colored images and stereo depth maps. The optimization process is driven by minimizing the following loss functions:

$$L_C = ||M \odot (\hat{C}_{g_p^2} - C)||, \quad L_D = ||M \odot (\hat{D}_{g_p^2}^{-1} - D^{-1})||, \quad (9)$$

where $\hat{C}_{g_p^2}$, $\hat{D}_{g_p^2}$, C, and D denote the rendered image, rendered depth, ground-truth image, and stereo depth, respectively. The overall training loss L is the sum of L_C and L_D. During inference, the trained model Ψ^* can render new views of the surgical scene as follows:

$$I_j = \Psi^*(T_j, t_j; K), \quad (10)$$

where K is the camera intrinsic matrix, T_j denotes extrinsic matrix representing the camera viewpoint at the unseen j-th frame, and t_j is the timestamp. Our model generates the extrapolated view I_j based on these inputs.

IV. Experiments

A. Experiment Setting

Datasets and Metrics. We evaluate our EndoRD-GS method and compare it with existing works on three datasets: 1) **The StereoMIS dataset** [52] consists of stereo endoscopic videos captured from in vivo porcine subjects, featuring diverse anatomical structures and challenging scenes with significant tissue deformations. 2) **The SCARED dataset** [53] provides RGB-D images with ground truth depth information from five porcine cadaver abdominal anatomies. 3) **The EndoNeRF dataset** [16] contains six stereo endoscopic video clips extracted from Da Vinci robotic prostatectomy procedures. Specifically, we use two public camera-motion-free scenes from the EndoNeRF dataset, three camera-motion-free clips following [2] from videos $P2\_1$ and $P3$ in the StereoMIS dataset ($P2\_1$: [1, 247], $P3$: [9100, 9467], and [11000, 11400]), and the five ordinary clips from the SCARED dataset involving varied camera motions for our main comparisons. Following [2], we split each scene's frames into training and testing sets at a 7:1 ratio. We evaluate reconstruction performance using PSNR, SSIM, and LPIPS metrics to assess the similarity between actual and rendered RGB images.

Implementation Details. We process each endoscopic scene by normalizing the video duration to the [0, 1] interval and resizing the input frames to a resolution of 512×512 pixels. Following [2], we initialize the Gaussian point cloud using the MAPF method and employ 17 learnable Gaussian basis functions per Gaussian primitive. The training phase consists of 3,000 iterations for the EndoNeRF dataset and 6,000 for the SCARED and StereoMIS datasets, using an initial learning rate of 1.6×10^{-3}. All experiments use the PyTorch framework on a single NVIDIA RTX 3090 GPU.

B. Comparison with State-of-the-art Methods

Quantitative Results. Table I demonstrates our method's superior performance across all datasets. In the StereoMIS dataset, our approach achieves a PSNR of 33.17 dB, outperforming the second-best method, Endo4D-GS, by 0.48 dB and the third-best method, Deform3DGS, by 0.52 dB. Our SSIM score of 0.873 exceeds Deform3DGS by 1.1%. Additionally, our method achieves an LPIPS of 0.160, outperforming Deform3DGS by 0.025 and Endo4D-GS by a significant margin of 0.012. On the SCARED dataset, our method reaches a PSNR of 28.91 dB, which is higher than the second-best method, Deform3DGS, by 0.34 dB and surpasses the third-best method, LGS, by 1.86 dB. In terms of SSIM, our method achieves a score of 0.842, improving upon Deform3DGS by 0.7% and LGS by 1.6%. The LPIPS score of 0.202 ties with Deform3DGS for the lowest value and is significantly lower than the third-best method, EndoGaussian, by 0.07. On the EndoNeRF dataset, our approach achieves a PSNR of 38.50 dB, surpassing the second-best result (Deform3DGS) by 0.27 dB, and matches EndoGaussian's SSIM score of 0.963.

Qualitative Results. Fig. 6 demonstrates our method's reconstruction capabilities on the SCARED and StereoMIS datasets compared to other state-of-the-art approaches. Our method accurately captures fine-grained tissue textures, including subtle surface variations, blood vessel patterns, and specular highlights on moist tissue surfaces. The yellow arrows in the magnified views highlight our superior preservation of sharp tissue boundaries and detailed deformation patterns. Fig. 7 showcases results from the EndoNeRF dataset during 'pulling' and 'cutting' procedures. Our method accurately reconstructs both global prostate anatomy and local tissue details, preserving important surgical landmarks such as vascular structures, tissue layer boundaries, and surface texture variations. The zoomed-in regions demonstrate our method's ability to maintain clear tissue edges and capture subtle changes in tissue appearance during surgical manipulation.

In summary, these results highlight that our EndoRD-GS excels in capturing intricate local tissue deformations, as demonstrated on the selected EndoNeRF and StereoMIS sequences, and achieves superior reconstruction quality under varied camera movements in the SCARED dataset.

C. Ablation Study

We conducted comprehensive ablation studies to evaluate our two key technical innovations: the periodic modulated Gaussian functions and the Biplane module. Our robustness analysis further evaluates the performance of both techniques across various scenarios, including perspective transformations and intricate instrument-tissue interactions.

1) Effectiveness of Each Proposed Component: We conduct ablation experiments to assess the effectiveness of each critical component across three datasets and report results in Table II. First, we examine the StereoMIS dataset. Using only the Gaussian basis function $\tilde{b}$, the baseline model achieves a PSNR of 32.65 dB and an SSIM of 0.8610, with a training time of 126.3 seconds. Incorporating the original HexPlane into the $\tilde{b}$ yields a modest improvement, with PSNR increasing to 32.76 dB and SSIM to 0.8630, but at the cost of significantly increased training time (279.0 seconds). When replacing HexPlane with our Biplane module while maintaining the $\tilde{b}$, we achieve better performance (PSNR: 32.79 dB, SSIM: 0.8637) despite reduced computational complexity. This counterintuitive improvement in performance can be attributed to several factors. First, the

TABLE I: Quantitative comparison of our EndoRD-GS with various methods across different datasets. yellow indicates the third-best result, light orange indicates the second-best result, and dark orange highlights the best result among these methods.

Method	StereoMIS [52]			SCARED [53]			EndoNeRF [16]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
EndoNeRF [16] (MICCAI 2022)	28.86	0.741	0.270	24.35	0.768	0.397	36.06	0.933	0.089
EndoSurf [17] (MICCAI 2023)	29.87	0.809	0.303	25.02	0.802	0.356	36.53	0.954	0.074
ForPlane-32k [42] (TMI 2024)	30.35	0.783	0.301	23.59	0.762	0.348	36.65	0.947	0.056
Endo4D-GS [51] (MICCAI 2024)	32.69	0.850	0.172	24.50	0.786	0.294	37.20	0.957	0.037
EndoGS [22] (MICCAI 2024)	29.81	0.711	0.346	26.46	0.770	0.339	36.84	0.963	0.041
EndoGaussian [23] (MICCAI 2024)	30.25	0.828	0.210	26.89	0.825	0.272	37.84	0.963	0.054
LGS [25] (MICCAI 2024)	31.36	0.816	0.231	27.05	0.826	0.297	37.48	0.955	0.068
Deform3DGS [2] (MICCAI 2024)	32.65	0.862	0.185	28.57	0.835	0.212	38.23	0.961	0.053
Ours: EndoRD-GS	**33.17**	**0.873**	**0.160**	**28.91**	**0.842**	**0.202**	**38.50**	**0.963**	**0.048**

TABLE II: Quantitative comparison of various component combinations across different datasets. Results are presented as mean ± standard deviation. The purple row indicates results using only the Gaussian basis function $\tilde{b}$. The row highlighted in orange shows the results from the Gaussian basis function in conjunction with HexPlane (Hex). The gray row presents results using our proposed periodic modulated Gaussian functions $\tilde{b}^*$ and the Biplane module (BIP).

Method Components				StereoMIS [52]			SCARED [53]			EndoNeRF [16]		
$\tilde{b}$	Hex	BIP	$\tilde{b}^*$	PSNR ↑	SSIM ↑	Training (s) ↓	PSNR ↑	SSIM ↑	Training (s) ↓	PSNR ↑	SSIM ↑	Training (s) ↓
✓				32.65 ±1.03	0.8610 ±0.03	126.3 ±3.2	28.57 ±4.53	0.8356 ±0.11	234.8 ±15.6	38.23 ±0.19	0.9615 ±0.0025	50.0 ±1.0
✓	✓			32.76 ±0.96	0.8630 ±0.03	279.0 ±11.0	28.82 ±4.42	0.8406 ±0.11	418.4 ±12.9	38.55 ±0.13	0.9630 ±0.0014	95.0 ±2.0
✓		✓		32.79 ±0.83	0.8637 ±0.02	174.3 ±5.9	28.90 ±4.53	0.8420 ±0.11	282.6 ±17.2	38.28 ±0.16	0.9620 ±0.0021	71.5 ±1.5
			✓	32.99 ±1.21	0.8707 ±0.03	133.7 ±5.7	28.58 ±4.50	0.8366 ±0.11	241.5 ±16.7	38.47 ±0.41	0.9630 ±0.0028	53.0 ±2.0
		✓	✓	**33.17 ±1.08**	**0.8733 ±0.03**	**182.7 ±4.2**	**28.91 ±4.39**	**0.8422 ±0.11**	**292.0 ±18.8**	**38.50 ±0.34**	**0.9635 ±0.0032**	**74.0 ±2.0**

TABLE III: Quantitative comparison between the baseline ($\tilde{b}$) and our Biplane module on 27 camera motion sequences.

Method		Seq. 1		Seq. 2		Seq. 3		Seq. 4		Seq. 5		Seq. 6		Seq. 7		Seq. 8		Seq. 9	
$\tilde{b}$	BIP	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓		26.58	66.1%	31.22	79.6%	30.46	79.6%	31.84	80.4%	26.46	75.8%	25.15	65.8%	37.33	92.9%	30.12	86.1%	31.68	89.8%
✓	✓	26.13	63.0%	30.19	75.7%	29.93	76.6%	30.76	76.2%	25.73	71.6%	24.54	60.9%	36.82	92.2%	29.66	85.0%	30.69	87.1%
Δ ↑		**+0.45**	**+3.1**	**+1.03**	**+3.9**	**+0.53**	**+3.0**	**+1.08**	**+4.2**	**+0.73**	**+4.3**	**+0.61**	**+4.9**	**+0.51**	**+0.7**	**+0.46**	**+1.1**	**+0.99**	**+2.7**

Method		Seq. 10		Seq. 11		Seq. 12		Seq. 13		Seq. 14		Seq. 15		Seq. 16		Seq. 17		Seq. 18	
$\tilde{b}$	BIP	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓		28.82	75.3%	25.65	65.8%	32.76	82.5%	35.92	93.6%	27.35	70.0%	33.91	89.6%	25.56	69.5%	33.01	84.9%	31.20	77.8%
✓	✓	28.21	71.0%	25.18	62.7%	32.10	80.8%	34.10	91.0%	26.73	66.1%	32.16	86.3%	25.07	67.0%	32.06	81.7%	30.31	74.0%
Δ ↑		**+0.61**	**+4.3**	**+0.47**	**+3.1**	**+0.66**	**+1.7**	**+1.82**	**+2.6**	**+0.62**	**+3.9**	**+1.75**	**+3.3**	**+0.49**	**+2.5**	**+0.95**	**+3.3**	**+0.88**	**+3.8**

Method		Seq. 19		Seq. 20		Seq. 21		Seq. 22		Seq. 23		Seq. 24		Seq. 25		Seq. 26		Seq. 27	
$\tilde{b}$	BIP	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓		36.41	93.3%	36.97	93.0%	27.00	77.3%	23.75	56.9%	23.63	53.3%	27.07	77.2%	27.99	82.6%	22.70	54.5%	23.39	61.6%
✓	✓	34.41	90.7%	35.87	91.7%	26.36	74.7%	23.19	52.3%	23.15	49.0%	26.60	74.7%	27.58	81.2%	22.28	51.0%	23.00	58.2%
Δ ↑		**+2.00**	**+2.6**	**+1.10**	**+1.3**	**+0.64**	**+2.6**	**+0.56**	**+4.6**	**+0.48**	**+4.3**	**+0.47**	**+2.6**	**+0.41**	**+1.3**	**+0.41**	**+3.5**	**+0.38**	**+3.4**

HexPlane component serves as a second-stage refinement that adjusts the initial deformations modeled by $\tilde{b}$. Given that endoscopic scenes exhibit confined and uniform structures, redundant grid parameters at the refinement stage could introduce unnecessary complexity and potential noise. Second, to align with the rapid deployment requirements of surgical applications, both configurations are trained for 6,000 iterations. In this limited training regime, the efficient grid design of our Biplane module demonstrates faster convergence with notably reduced training time (174.3 seconds), while the parameter-heavy HexPlane variant struggles to converge effectively yet fails to achieve comparable quality.

Besides, introducing only our periodic modulated Gaussian function $\tilde{b}^*$ leads to a more substantial performance boost. The PSNR increases to 32.99 dB and SSIM to 0.8707, with a training time of 133.7 seconds that remains comparable to the baseline. These results demonstrate that the periodic modulated Gaussian function alone significantly improves performance without additional computational cost. Finally, combining our periodic modulated Gaussian function and the Biplane module achieves the best performance. The PSNR reaches 33.17 dB, and the SSIM improves to 0.8733, with a training time of 182.7 seconds. This combination outperforms all previous configurations, confirming the synergistic effect of integrating both proposed components.

Similar trends are observed on the SCARED and EndoNeRF datasets, where the complete EndoRD-GS consistently outperforms other methods in terms of PSNR and SSIM met-

Fig. 6: Qualitative comparisons with state-of-the-art methods on the SCARED and StereoMIS datasets, highlighted in green and blue boxes. The details indicated by yellow arrows demonstrate that our method accurately reconstructs scene details. **T** denotes the normalized frame time $T \in [0, 1]$ (0 = first frame, 1 = last frame), proportional to actual time but not measured in seconds.

rics while maintaining training times within practical limits, demonstrating the superiority of each proposed component.

2) Robustness Analysis under Extreme Scenarios: We conduct experiments to evaluate our method's performance in handling intricate local tissue deformations and drastic perspective transformation scenarios.

To evaluate the impact of our periodic modulated Gaussian functions (PM) on simulating complex tissue deformations, we select three extremely intricate instrument-tissue interaction scenes: a cutting scene from the EndoNeRF dataset and two specific scenes from the StereoMis dataset P3: [9100, 9467], [11000, 11400]. We compare the baseline Deform3DGS model with our PM across these three complex instrument-tissue interactions. As shown in Fig. 8, our PM improves PSNR by 1.2%, 1.4%, and 1.1% in the cutting scene and sequences 1 and 2, respectively, demonstrating the effectiveness of our modulation schemes in capturing intricate tissue deformations. To visually demonstrate effectiveness, we further present a qualitative comparison of our PM against previous GS-based methods on these challenging instrument-tissue interaction scenes in Fig. 9. The zoomed-in views highlight regions undergoing significant deformation due to surgical tool manipulation. The visualization shows that our PM component captures these complex deformations with superior textural fidelity, whereas other methods tend to exhibit blurred textures in these high-deformation areas.

To assess the effectiveness of our Biplane module (BIP) in accommodating rapid perspective transformations, we define three typical drastic transformations based on the camera parameters of datasets: (1) large-scale movement (LM), where the camera moves over 10 mm within 5 seconds; (2) large-angle rotation (LR), where camera rotation exceeds 0.4 degrees within ten frames, and (3) sudden motion (SM), defined as the rapid position or angle changes of the camera between adjacent frames. Based on these definitions, we select three scenes from the Scared dataset (d3k1, d6k1, d7k1) to represent the three drastic transformation scenarios mentioned above. Moreover, to further verify the effectiveness of BIP under diverse camera motion conditions, we curate an expanded set of challenging sequences. First, from the original StereoMIS dataset [52], we identify 20 sequences ($Seq.1$-$Seq, 20$) characterized by translational and rotational camera movements (translation

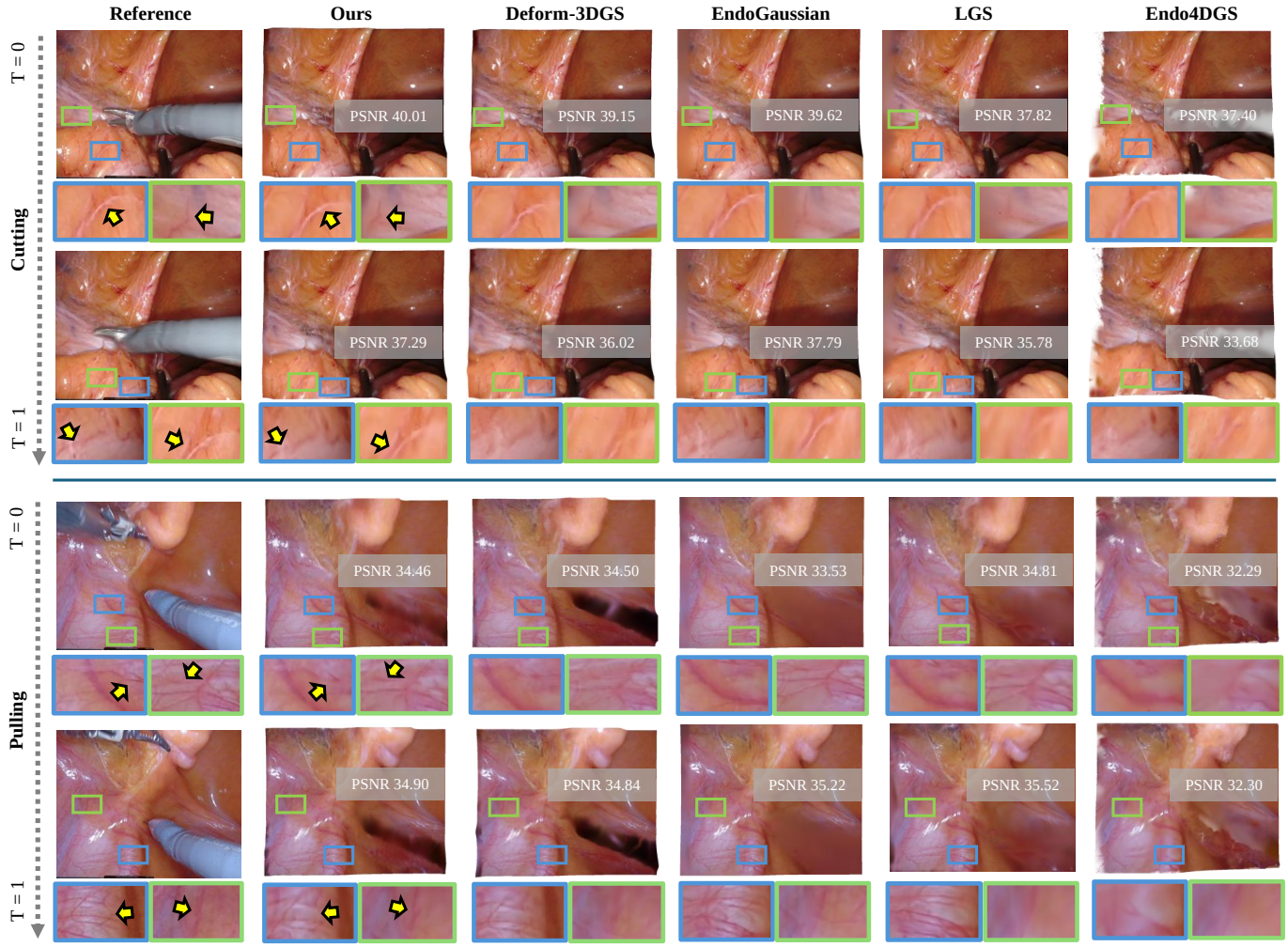


Fig. 7: Qualitative comparisons with state-of-the-art methods during the Cutting and Pulling phases of the EndoNeRF dataset, highlighted in green and blue boxes. Yellow arrows highlight tissue deformation details accurately reconstructed by our method.**T** denotes the normalized frame time $T \in [0, 1]$ (0 = first frame, 1 = last frame), proportional to actual time but not measured in seconds.

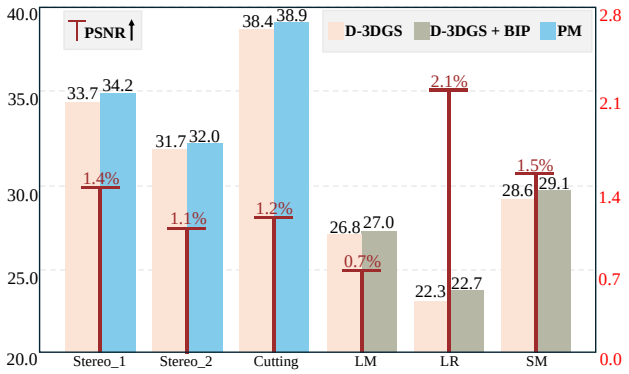


Fig. 8: Comparisons of Deform3DGS (D-3DGS) with our periodic modulated Gaussian functions (PM) in tissue-instrument interaction scenes and with our Biplane module (D-3DGS + BIP) in perspective transformation scenes. PSNR improvements are shown in red.

threshold ≥ 0.01 m, rotation threshold $\geq 0.05°$). Complementing these, we utilize the same standard to identify seven further sequences ($Seq.21$-$Seq.27$) exhibiting substantial camera jitter from the StereoMIS motion dataset recently released in [54]. This procedure selects a set of 27 StereoMIS sequences for evaluating the BIP in camera-motion scenarios. Detailed camera motion statistics for these sequences are provided in Table V.

We first conduct experiments comparing the baseline model with and without the Biplane module under perspective transformation conditions in the SCARED dataset. Numerical results in Fig. 8 demonstrate that the Biplane module consistently outperforms the baseline model, improving PSNR by 0.7%, 2.1%, and 1.5% in LM, LR, and SM scenarios, respectively. The qualitative comparison against previous GS-based methods under three types of perspective transformations demonstrates that our BIP effectively preserves scene consistency and structural details under rapid viewpoint changes, with clear improvements visible in the zoomed-

Fig. 9: Visualizations of reconstructed scenes under three extreme perspective transformation scenarios (upper part) and instrument-tissue interaction scenarios (lower part). Integrating our Biplane module with Deform3DGS (D-3DGS) yields more accurate detail reconstruction across comparisons with previous GS-based methods, as highlighted in blue boxes. Meanwhile, our proposed periodic modulated Gaussian functions (PM) achieve the most detailed tissue deformation reconstructions against previous GS-based methods, as highlighted in green boxes. **T** denotes the normalized frame time $T \in [0, 1]$ (0 = first frame, 1 = last frame), proportional to actual time but not measured in seconds.

in regions, as displayed in Fig. 9. We then conduct the experiments for our BIP on the new 27 challenging StereoMIS sequences and report the results in Table III. Across all camera motion sequences, the inclusion of BIP consistently enhances reconstruction quality over the baseline ($\tilde{b}$ only). Specifically, BIP delivers improvements across all 27 sequences, with gains frequently ranging between 0.4 and 0.9 dB in PSNR, and in several instances, such as $Seq.15$, exceeding 1.75 dB. Similarly, SSIM scores are enhanced in every sequence, with gains frequently ranging between 2.0% to 4.0%, with $Seq.6$ achieving an improvement over 4.9%. These results demonstrate that our BIP effectively improves reconstruction fidelity under a diverse range of camera motions.

Moreover, to evaluate the robustness and consistency of our full model, we compare the average performance metrics (PSNR, SSIM, and LPIPS) of prior methods under both standard scenarios and extreme scenarios, as detailed in Table IV. The extreme scenarios comprise the previously defined three instrument-tissue interaction scenes and three drastic perspective transformation scenes. Under standard scenarios, our EndoRD-GS outperforms other methods across all metrics, achieving a PSNR of 34.18 dB compared to Deform3DGS's 33.94 dB, an SSIM of 0.900 versus 0.895, and the lowest LPIPS score of 0.109 against 0.112, indicating superior visual fidelity. In extreme scenarios, while all methods experience a decline in performance, our EndoRD-GS demonstrates the

TABLE IV: Comparisons with state-of-the-art methods under standard and extreme scenarios. Performance drops are indicated in red, demonstrating our method's robust performance.

Standard Scenarios			
Method	PSNR ↑	SSIM ↑	LPIPS ↓
Endo4D-GS [51]	32.92	0.873	0.131
LGS [25]	32.72	0.874	0.161
EndoGaussian [23]	33.47	0.870	0.154
Deform3DGS [2]	33.94	0.895	0.112
EndoRD-GS (Ours)	**34.18**	**0.900**	**0.109**

Extreme Scenarios			
Method	PSNR ↑	SSIM ↑	LPIPS ↓
Endo4D-GS [51]	29.47 (-10.5%)	0.838 (-4.0%)	0.225 (+71.8%)
LGS [25]	28.29 (-13.5%)	0.821 (-6.1%)	0.285 (+77.0%)
EndoGaussian [23]	28.03 (-16.3%)	0.829 (-4.7%)	0.264 (+71.4%)
Deform3DGS [2]	30.25 (-10.9%)	0.851 (-4.9%)	0.176 (+57.1%)
EndoRD-GS (Ours)	**30.72** (-10.1%)	**0.860** (-4.4%)	**0.170** (+56.0%)

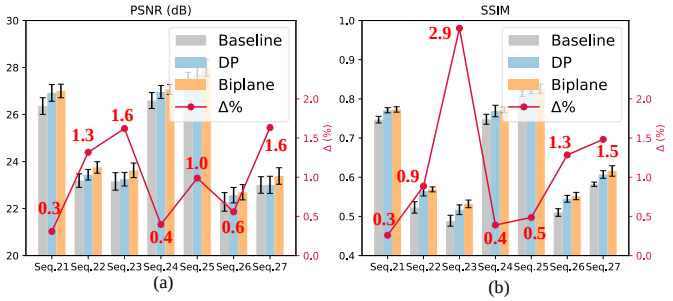


Fig. 10: Numerical PSNR and SSIM comparisons of Biplane against Double-Plane (DP) and Deform-3DGS (Baseline) on large jitter set sequences [54] ($Seq.21-27$).

highest consistency. It maintains superior performance and exhibits the smallest percentage decreases across PSNR and SSIM metrics. Specifically, our method's PSNR decreases by 10.1% to 30.72 dB, the most minor drop compared to EndoGaussian's 16.3% reduction to 28.03 dB and Deform3DGS's 10.9% decrease to 30.25 dB. Similarly, the SSIM decreases by only 4.4% to 0.860, less than Deform3DGS's 4.9% drop to 0.851 and LGS's 6.1% reduction to 0.821. Although LPIPS scores increase under challenging conditions, our method maintains the lowest score of 0.170, demonstrating that our EndoRD-GS is more resilient under challenging conditions. These results confirm that our EndoRD-GS, with our proposed periodic modulated Gaussian functions and the Biplane module, effectively modulates intricate tissue deformations and adapts to challenging perspective transformations, ensuring superior performance and consistency across various surgical scenes.

V. Discussion

A. Advantage Analysis of the Biplane Module

Comparison against Double-Plane. To rigorously analyze the advantages of our Biplane module, we compare it against the Double-Plane (DP), a simplified variant of the original HexPlane [1], retaining only its XY spatial and ZT spatiotemporal planes. While both DP and our Biplane ultimately combine features from two planes, a critical architectural divergence lies in how these features are extracted and interacted. In the DP configuration, it preserves only the XY and ZT grids and assigns them two independent groups of D channels. After sampling, it produces two vectors $g_{xy} \in \mathbb{R}^D$ and $h_{zt} \in \mathbb{R}^D$, which are not inherently constrained to an identical, pre-aligned feature space to represent the rest cross-dimensional couplings (XZ, XT, YZ, YT). These cross-dimensional couplings are therefore absent at this stage and can only be represented by subsequent tiny MLPs, which might not optimally model rich spatial-temporal features. In contrast, our Biplane explicitly learns both grids within one shared identical C-dimensional feature space. Channel-wise product of these grids then becomes:

$$\begin{aligned} f_c(x,y,z,t) &= g_{s,c}(x,y)\, h_{t,c}(z,t) = (a_1 x + a_2 y)(b_1 z + b_2 t) \\ &= a_1 b_1\, xz + a_1 b_2\, xt + a_2 b_1\, yz + a_2 b_2\, yt, \end{aligned} \tag{11}$$

where the coefficients a_1, a_2, b_1, b_2 are learnable parameters. This desgin recovers all four crucial cross-dimensional inter-

TABLE V: Camera-motion statistics. Upper block: 20 motion sequences from [52]; lower block: 7 large-jitter sequences from [54]. Translation expressed in centimeters.

Motion Set [52]							
Seq.	Translation (cm)			Rotation (deg)			Frame range
	Total	Net	Max	Total	Net	Max	
1 (P2_4)	0.9	0.6	0.1	9.173	7.527	0.654	[5669, 5741]
2 (P2_2)	1.9	1.9	0.1	3.272	3.049	0.223	[3535, 3593]
3 (P2_2)	1.8	1.8	0.1	6.982	6.900	0.451	[1691, 1787]
4 (P2_2)	0.7	0.5	0.1	6.238	5.164	0.551	[3323, 3391]
5 (P2_2)	0.9	0.8	0.1	5.894	5.849	0.265	[1209, 1331]
6 (P2_2)	1.3	1.3	0.1	6.997	6.895	0.469	[231, 353]
7 (P1)	0.4	0.4	0.1	5.914	5.903	0.394	[2377, 2455]
8 (P1)	1.8	1.5	0.1	26.150	23.644	1.148	[16265, 16383]
9 (P2_4)	1.8	1.7	0.5	6.924	6.850	1.800	[3771, 3847]
10 (P2_1)	4.9	4.8	0.2	14.919	14.204	0.591	[3043, 3187]
11 (P2_4)	2.2	2.1	0.1	7.782	7.563	0.305	[3075, 3185]
12 (P2_2)	0.9	0.8	0.1	5.761	5.214	0.315	[1935, 2017]
13 (P2_0)	0.7	0.7	0.1	8.756	8.126	0.451	[11059, 11143]
14 (P2_2)	1.4	1.1	0.1	9.173	8.336	0.667	[129, 225]
15 (P1)	0.8	0.6	0.1	27.271	24.061	1.853	[15649, 15745]
16 (P2_4)	2.6	2.5	0.2	7.884	7.633	0.897	[3189, 3289]
17 (P2_2)	3.3	3.3	0.3	7.945	5.900	0.790	[5005, 5083]
18 (P2_2)	2.3	2.3	0.1	4.248	2.894	0.212	[5087, 5175]
19 (P2_0)	0.9	0.8	0.1	10.853	9.538	0.658	[10969, 11051]
20 (P1)	0.8	0.7	0.1	9.934	9.219	0.739	[845, 915]
Large-Jitter Set [54]							
Seq.	Translation (cm)			Rotation (deg)			Frame range
	Total	Net	Max	Total	Net	Max	
21 (P2_6)	13.8	7.6	0.7	30.189	16.661	2.727	[10528, 10645]
22 (P2_1)	6.1	5.0	0.2	39.198	29.405	1.716	[5977, 6277]
23 (P2_2)	3.8	1.9	0.3	19.379	12.920	1.957	[49, 229]
24 (P2_6)	10.2	6.3	0.7	38.708	30.634	3.229	[10438, 10525]
25 (P2_6)	2.6	2.1	0.2	20.937	16.929	1.209	[10951, 11104]
26 (P2_1)	5.5	4.3	0.3	35.633	26.317	2.057	[6775, 7027]
27 (P2_2)	5.5	5.3	0.2	35.892	34.392	1.186	[235, 517]

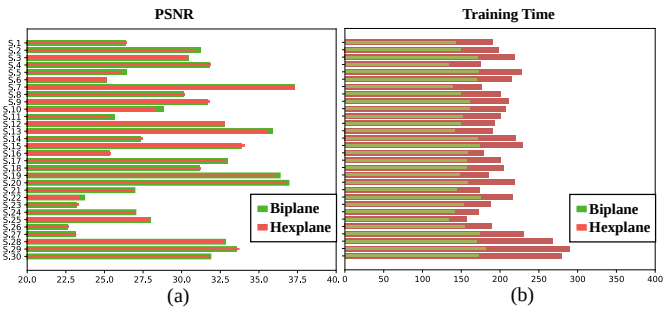


Fig. 11: Comparisons of PSNR and training time between our Biplane module and HexPlane on 30 sequences from the StereoMIS dataset.

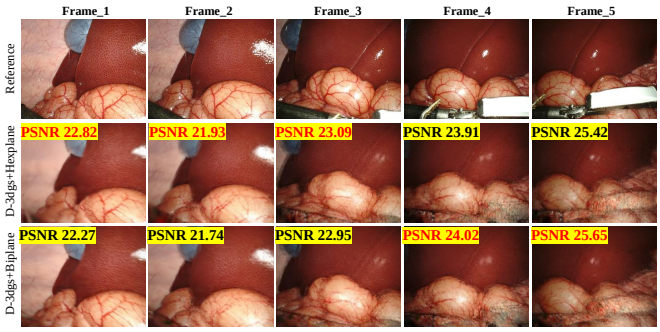


Fig. 12: Qualitative comparisons between our Biplane module and HexPlane on motion frames.

action terms for each channel $c \in C$, and allows our Biplane module to achieve a richer and direct representation of these spatiotemporal plane interactions.

To empirically validate the architectural benefits of the Biplane module, we implement the Double-Plane configuration based on the ablation design of the HexPlane and conduct direct comparisons with both our Biplane module and the Deform3DGS baseline. Quantitative results on the challenging large jitter sequences (Seq. 21-27 from the camera motion dataset [54]) are detailed in Fig. 10. Notably, our Biplane module consistently outperforms the Double-Plane, yielding an average improvement of 0.97 dB in PSNR and 1.1% in SSIM, with gains reaching up to 1.6 dB in PSNR and 2.9% in SSIM on the $Seq.23$. These findings underscore the advantages of our Biplane over the Double-Plane configuration.

Evaluation on Reconstruction Fidelity and Efficiency: Biplane vs. HexPlane. In addition to the performance comparison between our Biplane module and the HexPlane in Table II, we further conduct a comprehensive exploration to investigate the trade-offs between reconstruction fidelity and computational efficiency for our Biplane and HexPlane. Specifically, we evaluate both methods on a set of 30 sequences: three camera-motion-free sequences ($P2\_1$: [1, 247], $P3$: [9100, 9467], and [11000, 11400]; here we denote them as $Seq.\ 28-30.$), 20 camera-motion sequences ($Seq.\ 1-20$) from the original StereoMIS dataset [52], and seven sequences ($Seq.\ 21-27$) from the motion-augmented StereoMIS dataset [54], with results presented in Fig. 11. The quantitative results in Fig. 11(a) reveals that our Biplane frequently matches or surpasses HexPlane's performance. However, HexPlane occasionally shows marginal gains on a few specific sequences, such as $Seq.\ 23$ and $Seq.\ 27$. To better clarify the cases in which the HexPlane exhibits marginal gains, we provide additional qualitative examples, presenting frame-by-frame comparisons between our Biplane module and the HexPlane. As illustrated in Fig. 12, under conditions of multi-axial camera motions (exemplified by frames 1-3), the HexPlane might occasionally yield marginally higher numerical PSNR values, benefiting from its comprehensive use of multiple orthogonal planes. Other than these cases, our Biplane could achieve competitive and even superior performances compared to the HexPlane. Notably, Fig. 11(b) highlights that our Biplane offers a substantial reduction in training time compared to HexPlane, underscoring our Biplane's superior computational efficiency. This comprehensive exploration demonstrates that our Biplane module achieves a balance in high-fidelity reconstructions and training efficiency.

TABLE VI: Inference speed (FPS) under sequences with static and dynamic camera scenarios.

Method	Average FPS ± std	
	Static camera	Dynamic camera
Biplane	**199.68** ± 5.82	**154.27** ± 16.74
Double-Plane	183.99 ± 6.13	147.59 ± 12.28
HexPlane	151.70 ± 4.78	120.14 ± 8.43

TABLE VII: Mean PSNR (↑) comparison of GBF and GBF+Biplane on the StereoMIS training set under different training iterations.

Method	6000 iters	10000 iters
GBF + Biplane	34.11	**35.05**
GBF	**34.24**	35.01

Inference Time Comparison: Biplane vs. Double-Plane and HexPlane. To further analyze the advantages of our proposed Biplane compared to the Double-Plane and Hexplane, we evaluate inference time in terms of frames per second (FPS) under two representative conditions: (1) static-camera scenarios using the pulling and cutting sequences from EndoNeRF [16], and (2) dynamic-camera scenarios with significant motion using seven StereoMIS sequences ($Seq.21-27$) [54]. As shown in Table VI, compared to Double-Plane (183.99/147.59 FPS) and HexPlane (151.70/120.14 FPS), our Biplane module consistently achieves faster rendering speed under both conditions, achieving 199.68 FPS for static cameras and 154.27 FPS for dynamic cameras. This acceleration may stem from our shared-feature-space design. Specifically, unlike Double-Plane, which maintains separate XY and ZT feature spaces, our Biplane learns both feature spaces within a unified C-dimensional basis. During inference, our design enables direct recovery of all cross-dimensional interactions (XZ, XT, YZ, YT) through channel-wise multiplication, reducing both subsequent MLP computational burden and memory access overhead. Moreover, by using only two grids instead of six grids in Hexplane, our Biplane achieves comparable representation capacity with much lower inference time.

Impact of the Biplane Module on Training Performance.

To further investigate the interplay between the Biplane and Gaussian Basis Functions (GBF), we evaluate the training performance on selected StereoMIS sequences ($Seq.28$–30), with results in Table VII. The results reveal that while the standard GBF model achieves slightly higher PSNR on the training set with 6k iterations, the addition of our Biplane module leads to consistently superior performance on the testing set, as reported in Table II. This phenomenon reflects two key factors. **First**, the pure GBF parameterization optimizes only two learnable parameters (the scale σ and center θ), allowing it to converge rapidly within fewer training iterations and achieve slightly higher PSNR on the training set. In contrast, the incorporation of the Biplane module introduces extra parameters, requiring more training iterations to converge, resulting in marginally lower PSNR on the training set. We further conduct experiments with 10k iterations on GBF versus GBF + Biplane, and summarize the performance on different iterations in Table VII. The results demonstrate that our Biplane could converge and perform better than GBF with additional training iterations. **Second**, our Biplane module imposes implicit regularization by enforcing weight sharing between the spatial (x, y) and spatiotemporal (z, t) feature grids through a unified C-dimensional feature space. Moreover, the incorporation of our Biplane module serves as an additional stage to model the deformation trajectories of Gaussian primitives. This enhanced modeling capacity allows our method to better generalize to unseen views, effectively handling complex primitive deformations caused by camera motion and local tissue deformation.

VI. Conclusion and Future Work

In this paper, we introduce EndoRD-GS, an advanced method for robust dynamic endoscopic scene reconstruction. EndoRD-GS comprises periodically modulated Gaussian functions for intricate tissue deformation modeling and a Biplane module that facilitates global adjustments for robust reconstruction during dramatic perspective transformations. Experiments on three endoscopic datasets demonstrate the superiority of our EndoRD-GS, with ablation studies confirming the effectiveness of each component.

Our approach aligns with leading contemporaneous works, such as Deform3DGS [2] and EndoGaussian [23]. The *real-time* capability is predominantly defined by the ability to render novel views at interactive frame rates. Despite the real-time rendering speed, the per-scene training duration remains a bottleneck for achieving truly instantaneous 3D reconstruction. To solve this, one promising direction involves developing generalizable foundation models [55]–[57] capable of zero-shot adaptation to new surgical scenes, thereby minimizing per-scene training requirements. Our future work will explore such generalizable 3D Gaussian representations to enable instant, high-fidelity dynamic 3D reconstruction for intra-operative workflows.

References

[1] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2023, pp. 130–141.

[2] S. Yang, Q. Li, D. Shen, B. Gong, Q. Dou, and Y. Jin, "Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting," *arXiv preprint arXiv:2405.17835*, 2024.

[3] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery," *IEEE RAL*, vol. 3, no. 1, pp. 155–162, 2017.

[4] M. Lerotic, A. J. Chung, J. Clark, S. Valibeik, and G.-Z. Yang, "Dynamic view expansion for enhanced navigation in natural orifice transluminal endoscopic surgery," in *MICCAI*. Springer, 2008, pp. 467–475.

[5] E. Pelanis, A. Teatini, B. Eigl, A. Regensburger, A. Alzaga, R. P. Kumar, T. Rudolph, D. L. Aghayan, C. Riediger, N. Kvarnström *et al.*, "Evaluation of a novel navigation platform for laparoscopic liver surgery with organ deformation compensation using injected fiducials," *Medical image analysis*, vol. 69, p. 101946, 2021.

[6] M. Ribeiro, Y. Espinel, N. Rabbani, B. Pereira, A. Bartoli, and E. Buc, "Augmented reality guided laparoscopic liver resection: a phantom study with intraparenchymal tumors," *Journal of Surgical Research*, vol. 296, pp. 612–620, 2024.

[7] H. Saeidi, H. N. Le, J. D. Opfermann, S. Léonard, A. Kim, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous laparoscopic robotic suturing with a novel actuated suturing tool and 3d endoscope," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 1541–1547.

[8] S. Gong, Y. Long, K. Chen, J. Liu, Y. Xiao, A. Cheng, Z. Wang, and Q. Dou, "Self-supervised cyclic diffeomorphic mapping for soft tissue deformation recovery in robotic surgery scenes," *IEEE TMI*, 2024.

[9] P. Brandao, D. Psychogyios, E. Mazomenos, D. Stoyanov, and M. Janatka, "Hapnet: hierarchically aggregated pyramid network for real-time stereo matching," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 3, pp. 219–224, 2021.

[10] H. Luo, C. Wang, X. Duan, H. Liu, P. Wang, Q. Hu, and F. Jia, "Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images," *Computers in biology and medicine*, vol. 140, p. 105109, 2022.

[11] H. Zhou and J. Jagadeesan, "Real-time dense reconstruction of tissue surface from stereo optical video," *IEEE TMI*, vol. 39, no. 2, pp. 400–412, 2019.

[12] H. Zhou and J. Jayender, "Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos," in *MICCAI 2021*. Springer, 2021, pp. 331–340.

[13] Y. Long, Z. Li, C. H. Yee, C. F. Ng, R. H. Taylor, M. Unberath, and Q. Dou, "E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception," in *MICCAI 2021*. Springer, 2021, pp. 415–425.

[14] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1438–1447, 2019.

[15] V. M. Batlle, J. M. Montiel, P. Fua, and J. D. Tardós, "Lightneus: Neural surface reconstruction in endoscopy using illumination decline," in *MICCAI*. Springer, 2023, pp. 502–512.

[16] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *MICCAI*. Springer, 2022, pp. 431–441.

[17] R. Zha, X. Cheng, H. Li, M. Harandi, and Z. Ge, "Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos," in *MICCAI*. Springer, 2023, pp. 13–23.

[18] C. Yang, K. Wang, Y. Wang, X. Yang, and W. Shen, "Neural lerplane representations for fast 4d reconstruction of deformable tissues," in *MICCAI*. Springer, 2023, pp. 46–56.

[19] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[20] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *NIPS*, vol. 33, pp. 2492–2502, 2020.

[21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[22] L. Zhu, Z. Wang, Z. Jin, G. Lin, and L. Yu, "Deformable endoscopic tissues reconstruction with gaussian splatting," *arXiv preprint arXiv:2401.11535*, 2024.

[23] Y. Liu, C. Li, C. Yang, and Y. Yuan, "Endogaussian: Gaussian splatting for deformable surgical scene reconstruction," *arXiv preprint arXiv:2401.12561*, 2024.

[24] C. Li, B. Y. Feng, Y. Liu, H. Liu, C. Wang, W. Yu, and Y. Yuan, "Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting," *arXiv preprint arXiv:2407.01029*, 2024.

[25] H. Liu, Y. Liu, C. Li, W. Li, and Y. Yuan, "Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction," *arXiv preprint arXiv:2406.16073*, 2024.

[26] H. Zhao, X. Zhao, L. Zhu, W. Zheng, and Y. Xu, "Hfgs: 4d gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction," *arXiv preprint arXiv:2405.17872*, 2024.

[27] A. K. Golahmadi, D. Z. Khan, G. P. Mylonas, and H. J. Marcus, "Tool-tissue forces in surgery: A systematic review," *Annals of Medicine and Surgery*, vol. 65, p. 102268, 2021.

[28] M. Camara, E. Mayer, A. Darzi, and P. Pratt, "Soft tissue deformation for surgical simulation: a position-based dynamics approach," *CARS*, vol. 11, pp. 919–928, 2016.

[29] J. Shin, Y. Zhong, and C. Gu, "Real-time nonlinear characterization of soft tissue mechanical properties," *Journal of Sensors*, vol. 2020, no. 1, p. 9873410, 2020.

[30] A. Idkaidek and I. Jasiuk, "Toward high-speed 3d nonlinear soft tissue deformation simulations using abaqus software," *Journal of robotic surgery*, vol. 9, pp. 299–310, 2015.

[31] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based processing and analysis of endoscopic images and videos: A survey," *Multimedia Tools and Applications*, vol. 77, pp. 1323–1362, 2018.

[32] T. Teufel, H. Shu, R. D. Soberanis-Mukul, J. E. Mangulabnan, M. Sahu, S. S. Vedula, M. Ishii, G. Hager, R. H. Taylor, and M. Unberath, "Oneslam to map them all: a generalized approach to slam for monocular endoscopic imaging based on tracking any point," *CARS*, pp. 1–8, 2024.

[33] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. Ellerbee Bowden, "3d reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical optics express*, vol. 8, no. 4, pp. 2106–2123, 2017.

[34] M. Hu, G. Penney, M. Figl, P. Edwards, F. Bello, R. Casula, D. Rueckert, and D. Hawkes, "Reconstruction of a 3d surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes," *MedIA*, vol. 16, no. 3, pp. 597–611, 2012.

[35] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[36] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on robotics*, vol. 37, no. 1, pp. 291–303, 2020.

[37] J. J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J. D. Tardós, and J. M. Montiel, "Sd-defslam: Semi-direct monocular slam for deformable and intracorporeal scenes," in *ICRA*. IEEE, 2021, pp. 5170–5177.

[38] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[39] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021, pp. 10 318–10 327.

[40] Y.-L. Liu, C. Gao, A. Meuleman, H.-Y. Tseng, A. Saraf, C. Kim, Y.-Y. Chuang, J. Kopf, and J.-B. Huang, "Robust dynamic radiance fields," in *CVPR*, 2023, pp. 13–23.

[41] B. G. Gerats, J. M. Wolterink, S. P. Mol, and I. A. Broeders, "Neural fields for 3d tracking of anatomy and surgical instruments in monocular laparoscopic video clips," *arXiv preprint arXiv:2403.19265*, 2024.

[42] C. Yang, K. Wang, Y. Wang, Q. Dou, X. Yang, and W. Shen, "Efficient deformable tissue reconstruction via orthogonal neural plane," *IEEE Transactions on Medical Imaging*, 2024.

[43] S. Saha, S. Liu, S. Lin, J. Lu, and M. Yip, "Based: Bundle-adjusting surgical endoscopic dynamic video reconstruction using neural radiance fields," *arXiv preprint arXiv:2309.15329*, 2023.

[44] X. Sun, F. Wang, Z. Ma, and H. Su, "Dynamic surface reconstruction in robot-assisted minimally invasive surgery based on neural radiance fields," *CARS*, vol. 19, no. 3, pp. 519–530, 2024.

[45] Y. Wang, B. Gong, Y. Long, S. H. Fan, and Q. Dou, "Efficient endonerf reconstruction and its application for data-driven surgical simulation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, no. 5, pp. 821–829, 2024.

[46] Q. Gao, Q. Xu, Z. Cao, B. Mildenhall, W. Ma, L. Chen, D. Tang, and U. Neumann, "Gaussianflow: Splatting gaussian dynamics for 4d content creation," *arXiv preprint arXiv:2403.12365*, 2024.

[47] Z. Li, Z. Chen, Z. Li, and Y. Xu, "Spacetime gaussian feature splatting for real-time dynamic view synthesis," in *CVPR*, 2024, pp. 8508–8520.

[48] S. Hu, T. Hu, and Z. Liu, "Gauhuman: Articulated gaussian splatting from monocular human videos," in *CVPR*, 2024, pp. 20 418–20 431.

[49] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *CVPR*, 2024, pp. 20 310–20 320.

[50] M. Kocabas, J.-H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan, "Hugs: Human gaussian splats," in *CVPR*, 2024, pp. 505–515.

[51] Y. Huang, B. Cui, L. Bai, Z. Guo, M. Xu, and H. Ren, "Endo-4dgs: Distilling depth ranking for endoscopic monocular scene reconstruction with 4d gaussian splatting," *arXiv preprint arXiv:2401.16416*, 2024.

[52] M. Hayoz, C. Hahne, M. Gallardo, D. Candinas, T. Kurmann, M. Allan, and R. Sznitman, "Learning how to robustly estimate camera pose in endoscopic videos," *CARS*, vol. 18, no. 7, pp. 1185–1192, 2023.

[53] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[54] M. Hayoz, C. Hahne, T. Kurmann, M. Allan, G. Beldi, D. Candinas, P. Márquez-Neila, and R. Sznitman, "Online 3d reconstruction and dense tracking in endoscopic videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 444–454.

[55] S. Szymanowicz, E. Insafutdinov, C. Zheng, D. Campbell, J. F. Henriques, C. Rupprecht, and A. Vedaldi, "Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image," *arXiv preprint arXiv:2406.04343*, 2024.

[56] B. Smart, C. Zheng, I. Laina, and V. A. Prisacariu, "Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs," *arXiv preprint arXiv:2408.13912*, 2024.

[57] Z. Wu, H. Xu, G. Xu, P. Nie, Z. Yan, J. Zheng, L. Qu, M. Li, and L. Nie, "Textsplat: Text-guided semantic fusion for generalizable gaussian splatting," *arXiv preprint arXiv:2504.09588*, 2025.